Final Project

Duc Viet Mai

NUID: 002661113

ALY6010

Probability Theory and Introductory Statistics

Date: December 16, 2022

Abstract. To begin with, the data set of interest is about global living costs in nearly 5000 cities worldwide, put together from Numbeo's website. The objective of the analysis is to check for the relationship between chosen variables to observe how they impact living costs. Specifically, I will perform the methodology of hypothesis testing and regression analysis to visualize these relationships. And in the final results, I hope there is significant evidence to conclude that there is a strong relationship between these variables.

Keywords: living costs, hypothesis testing, regression analysis, relationships, variables

(Note: I changed the data set used in milestones 1 and 2 due to its inability to perform a meaningful linear regression model with the Professor's approval.)

1. Introduction

I divided the subjects of interest into two categories dependent and independent variables. The dependent includes the price of utilities (electricity, heating, water, etc.), preschool's costs, and the cost of an apartment inside the City Centre. In contrast, the independent contains the price of gasoline, average monthly salary (after tax), and the price of an apartment outside the City Centre. I chose these subjects because they are necessities for a citizen to maintain a healthy daily life. The lack of any of these subjects would cause a substantial decrease in that individual's living quality.

The first significant question asks about the relationship between the price of gasoline and housing utilities. The second one addresses the correlation between the average monthly salary and the preschool's costs. And finally, the third crucial question looks into the relationship between apartment prices in and outside the City Centre.

In this paper, the second section will show all materials of the data set and the statistical analysis methods. The third section will depict the results of the EDA (Exploratory Data Analysis), with a summary of answers for questions one to three. Next, the final part will be the discussion, in which I will talk about the limitations of the analysis and possible future work.

2. Materials & Methods

2.1. Dataset

To be more specific about my dataset, it contains nearly 5000 observations of 59 variables. From my perspective, I believe that the author has gathered the data from the Numbeo website, which involves keeping track of the global cost of living and maintaining the quality of life. After cleaning the missing values in the data set, I ended with nearly 2000 observations left. Most variables are about the prices of food, housing, appliances, etc., and we can visualize the living costs based on the increase or decrease in these prices.

(Sources: https://www.kaggle.com/datasets/mvieira101/global-cost-of-living)

2.2. Statistical Analysis

1. Hypothesis tests

   Dependent variables:

   Case 1: One sample z-test

   - Null hypothesis: The mean price of an apartment in the City Centre is equal 0.
     H0: $\mu = 0$
   - Alternative hypothesis: The mean price of an apartment in the City Centre is not equal 0.
     H1: $\mu \neq 0$

   Case 2: One sample z-test

   - Null hypothesis: The mean costs of preschool (per kid) is equal 0.
     H0: $\mu = 0$
   - Alternative hypothesis: The mean costs of preschool (per kid) is not equal 0.
     H1: $\mu \neq 0$

   Case 3: One sample z-test

- Null hypothesis: The mean price of housing utilities is equal 0.
  H0: $\mu = 0$
- Alternative hypothesis: The mean price of housing utilities is not equal 0.
  H1: $\mu \neq 0$

Independent variables:

Case 4: One sample z-test

- Null hypothesis: The mean price of gasoline (per liter) is equal 0.
  H0: $\mu = 0$
- Alternative hypothesis: The mean price of gasoline (per liter) is not equal 0.
  H1: $\mu \neq 0$

Case 5: One sample z-test

- Null hypothesis: The mean average monthly salary is equal 0.
  H0: $\mu = 0$
- Alternative hypothesis: The mean average monthly salary is not equal 0.
  H1: $\mu \neq 0$

Case 6: One sample z-test

- Null hypothesis: The mean price of an apartment outside of Centre is equal 0.
  H0: $\mu = 0$
- Alternative hypothesis: The mean price of an apartment outside of Centre is not equal 0.
  H1: $\mu \neq 0$

2. Correlation tests

Question 1

- Null hypothesis: The correlation between the price of gasoline and housing utilities is equal 0.
  H0: $\rho = 0$
- Alternative hypothesis: The correlation between the price of gasoline and housing utilities is not equal 0.
  H1: $\rho \neq 0$

Question 2

- Null hypothesis: The correlation between the average monthly salary and costs of preschool is equal 0.
  H0: $\rho = 0$
- Alternative hypothesis: The correlation between the average monthly salary and costs of preschool is not equal 0.

H1: $\rho \neq 0$

Question 3

- Null hypothesis: The correlation between the price of an apartment in and outside the City Centre is equal 0.
  H0: $\rho = 0$
- Alternative hypothesis: The correlation between the price of an apartment in and outside the City Centre is not equal 0.
  H1: $\rho \neq 0$

3. Linear regression models

Question 1

$$Equation\ 1: \hat{y} = 5.899 + 87.271 * (x33)$$

(Note: y – the housing utilities, x33 – the price of gasoline)

Question 2

$$Equation\ 2: \hat{y} = 56.095800 + 0.250174 * (x54)$$

(Note: y – the costs of preschool, x54 – the average monthly salary)

Question 3

$$Equation\ 3: \hat{y} = 33.300363 + 1.239149 * (x49)$$

(Note: y – the price of an apartment inside the City Centre, x49 – the price of an apartment outside the City Centre)

4. Statistical software used

- Software: R Programming Language
- Packages: tidyverse, ggplot2, MASS, BDSA, gmodels

3. Results

3.1. Exploratory Data Analysis (EDA)

	Mean	Standard deviation	Median	Min	Max
X33 (N = 1264)	1.419	0.499	1.325	0.030	2.760
X36 (N = 1264)	129.760	78.815	122.670	10.000	1063.630
X42 (N = 1264)	485.950	494.421	312.58	10.690	2903.610
X54 (N = 1264)	1718.240	1525.716	1114.010	35.750	7910.520
X48 (N = 1264)	701.810	675.092	514.480	43.810	12608.830
X49 (N = 1264)	539.490	529.734	389.510	22.280	8989.370

In the first question, I have x33 to represent the price of gasoline and x 36 as the costs of housing utilities. As noticeable, the mean is larger than the median in both variables, which indicates a right-skewed distribution. The low standard deviation value implies the data points variation is small, which means the gas and housing utility prices are similar globally.

In the second question, x42 is the cost of preschool, and x54 is the average monthly salary. And in the third question, x48 is an apartment's price inside the City Centre, and x49 is the price outside. As you can see, these variables have a heavily right-skewed distribution as their means are higher than medians. Furthermore, all of them have a high standard deviation value, which I believe is because of the existence of outliers in the data set. However, I chose not to remove them because some might contain vital information. For instance, the average salary can be higher in developed countries like US or China than in developing countries like Thailand or Vietnam. You could observe this as the min of x54 is 35, significantly smaller than its maximum of 7910.

3.2. The relationship between the price of gasoline and housing utilities

	Z-score ($\alpha = 0.05$)	P-value ($\alpha = 0.05$)
X33	101.100	2.2e-16
X36	58.532	2.2e-16

Case 3 results: The two-tailed test results show that our z-critical value is 101.100 and our p-value is 2.2e-16, smaller than the alpha value (0.05) for the 95% confidence level.

Therefore, we reject the null hypothesis. We might interpret this result as the mean price of housing utilities is not equal 0.

Case 4 results: The two-tailed test results show that our z-critical value is 58.532 and our p-value is 2.2e-16, smaller than the alpha value (0.05) for the 95% confidence level. Therefore, we reject the null hypothesis. We might interpret this result as the mean price of gasoline is not equal to 0.

	T-score (df = 1262)	**P-value (α = 0.05)**	**Correlation**	**R-squared**	**Residual standard error**
X33 & X36	23.556	2.2e-16	0.553	0.305	65.710

Question 1 results: The two-tailed test results show that our t-critical value is 23.556 and our p-value is 2.2e-16, smaller than the alpha value (0.05) for the 95% confidence level. Therefore, we reject the null hypothesis. We might interpret this result as the correlation between the price of gasoline and housing utilities is not equal to 0.

The correlation value is 0.553, which implies a strong and positive relationship between the price of gasoline in and utilities. Specifically, equation 1 above can further explain this as one unit increase in the price of gasoline associated with 87 unit increase in the costs of housing utilities. The logic is common sense as the price of fuel increases, electricity and heating costs will possibly increase relatively. The R-squared value is 0.305, indicating that the regression model can explain 30.5% of the predicted values in real life.

3.3. The relationship between the costs of preschool and average monthly salary

	Z-score (α = 0.05)	**P-value (α = 0.05)**
X54	40.039	2.2e-16
X42	34.944	2.2e-16

Case 2 results: The two-tailed test results show that our z-critical value is 34.944 and our p-value is 2.2e-16, smaller than the alpha value (0.05) for the 95% confidence level. Therefore, we reject the null hypothesis. We might interpret this result as the mean costs of preschool (per kid) is not equal 0.

Case 5 results: The two-tailed test results show that our z-critical value is 40.039 and our p-value is 2.2e-16, smaller than the alpha value (0.05) for the 95% confidence level. Therefore, we reject the null hypothesis. We might interpret this result as the mean average monthly salary is not equal 0.

	T-score (df = 1262)	P-value (α = 0.05)	Correlation	R-squared	Residual standard error
X54 & X42	43.147	2.2e-16	0.772	0.956	314.400

Question 2 results: The two-tailed test results show that our t-critical value is 43.147 and our p-value is 2.2e-16, smaller than the alpha value (0.05) for the 95% confidence level. Therefore, we reject the null hypothesis. We might interpret this result as the correlation between the average monthly salary and costs of preschool is not equal 0.

The correlation value is 0.772, showing a positive linear relationship between the average monthly salary and preschool costs. Specifically, equation 2 above can further explain this as one unit increase in the average salary associated with a rise of 0.25 units in the price of preschool. As the household's income increases, the school might decide to raise the tuition for further investment at school regarding the future benefits for their children. The R-squared value is 0.956, indicating that the regression model can explain 95.6% of the predicted values in real life, which means it is likely to forecast correctly the preschool costs based on monthly salary.

3.4. The relationship between the price of an apartment in and outside the City Centre

	Z-score (α = 0.05)	P-value (α = 0.05)
X49	36.207	2.2e-16
X48	36.960	2.2e-16

Case 1 results: The two-tailed test results show that our z-critical value is 36.960 and our p-value is 2.2e-16, smaller than the alpha value (0.05) for the 95% confidence level. Therefore, we reject the null hypothesis. We might interpret this result as the mean price of an apartment in the City Centre is not equal 0.

Case 6 results: The two-tailed test results show that our z-critical value is 36.207 and our p-value is 2.2e-16, smaller than the alpha value (0.05) for the 95% confidence level. Therefore, we reject the null hypothesis. We might interpret this result as the mean price of an apartment outside of Centre is not equal 0.

	T-score (df = 1262)	P-value (α = 0.05)	Correlation	R-squared	Residual standard error
X48 & X49	147.890	2.2e-16	0.972	0.945	157.700

Question 3 results: The two-tailed test results show that our t-critical value is 147.890 and our p-value is 2.2e-16, smaller than the alpha value (0.05) for the 95% confidence level. Therefore, we reject the null hypothesis. We might interpret this result as the correlation between the price of an apartment in and outside the City Centre is not equal 0.

The correlation value is 0.972 (nearly 1), indicating a perfect positive linear relationship between the apartment prices in and outside the City Centre. Specifically, equation 3 above can further explain this as one unit increase in the apartment price outside associated with a 1.24 unit increase in the price of apartments inside the City Centre. Undoubtedly, as the price of housing in the countryside increases, housing prices in the urban area must also increase, which might further create a housing bubble in the future. The R-squared value is 0.972, indicating that the regression model can explain 97.2% of the predicted values in real life. Henceforth, whenever you see the price of housing outside the Centre growing, you immediately know there is a surge in the housing price inside the Centre.

4. Discussion

4.1. This work

This paper aims to aware the international institutions in charge of maintaining high living standards for our society by monitoring the living costs, especially in developing countries with disadvantages in economic aspects, so that our future generations do not have to live in poverty and hunger. Firstly, by controlling the costs of fossil fuels and gas, we can save for housing utilities like electricity and heat for the winter. Next, the preschools need to deliberately raise the tuition fee based on the monthly household income because not any increase in the parent's salary means that they will suddenly be rich. By carefully supervising this, we can have a win-win situation as the children can still have a chance to have an education, and the school can still expand its facilities with adequate investment. Third and finally, the government should raise awareness of a potential housing bubble by looking into the relationship between housing prices in and outside the City Centre.

Overall, housing, education, and resources are the basic living costs people worldwide need to take care of daily to approach a decent life.

4.2. Limitations

The limitations of this paper include a not-so-large sample size due to the removal of many missing values. Many other variables that could affect the dependent variables are lacking, such as food prices or mortgage rates. Although we have decent values of R-squared, high values of residual standard errors still exist, implying a lot of data points variation.

4.3. Future work

The next step we can take in the future analysis is multiple regression with dummy variables representing countries, which we can observe the living costs behavior between them. Other questions I can ask are: Could mortgage rates affect housing prices? Would gasoline prices affect imported foods shipping from foreign countries? These paths will all seem to affect much on the living costs in the long run, which we should be mindful too.

5. References

[1] Bluman, A. G. (2018). *Elementary statistics: A step by step approach.* McGraw-Hill Education. 10th ed, 1- 368. ISBN: 978-1-259-75533-0.

[2] Kaggle: Your Home for Data Science. (n.d.). *Global cost of living.* Retrieve from: https://www.kaggle.com/datasets/mvieira101/global-cost-of-living

[3] Statistical tools for high-throughput data analysis (STHDA). *Ggplot2title: main, axis and legend titles.* Data Visualization. Retrieve from: http://www.sthda.com/english/wiki/ggplot2-title-main-axis-and-legend-titles

6. Appendix

Case 1:

```
        One-sample z-Test

data:  df$x48
z = 36.96, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 664.5893 739.0226
sample estimates:
mean of x
  701.806
```

```
[1] "Reject the null hypothesis. We have enough statistical evidence to conclude the mean price of apar
tment in City Centre is different from 0."
```

Case 2:

```
        One-sample z-Test

data:  df$x42
z = 34.944, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 458.6977 513.2108
sample estimates:
mean of x
 485.9542
```

```
[1] "Reject the null hypothesis. We have enough statistical evidence to conclude the mean price of kind
ergarden is different from 0."
```

Case 3:

```
        One-sample z-Test

data:  df$x36
z = 58.532, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 125.4104 134.1003
sample estimates:
mean of x
 129.7553
```

```
[1] "Reject the null hypothesis. We have enough statistical evidence to conclude the mean price of basi
c utilities is different from 0."
```

Case 4:

```
        One-sample z-Test

data:  df$x33
z = 101.1, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 1.391695 1.446723
sample estimates:
mean of x
 1.419209
```

```
[1] "Reject the null hypothesis. We have enough statistical evidence to conclude the mean price of gaso
line is different from 0."
```

Case 5:

```
        One-sample z-Test

data:  df$x54
z = 40.039, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 1634.127 1802.347
sample estimates:
mean of x
 1718.237
```

```
[1] "Reject the null hypothesis. We have enough statistical evidence to conclude the mean average month
ly salary is different from 0."
```

Case 6:

```
        One-sample z-Test

data:  df$x49
z = 36.207, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 510.2844 568.6911
sample estimates:
mean of x
 539.4877
```

```
[1] "Reject the null hypothesis. We have enough statistical evidence to conclude the mean price of apar
tment outside of Centre is different from 0."
```

EDA:

```
> summary(df$x33)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.030   1.130   1.325   1.419   1.860   2.760
> sd(df$x33)
[1] 0.4990913
> summary(df$x36)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.00   64.86  122.67  129.76  182.26 1063.63
> sd(df$x36)
[1] 78.81483
> summary(df$x42)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.69  160.94  312.58  485.95  630.14 2903.61
> sd(df$x42)
[1] 494.421
> summary(df$x54)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  35.75  497.35 1114.01 1718.24 2672.32 7910.52
> sd(df$x54)
[1] 1525.716
> summary(df$x48)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   43.81   280.63   514.48   701.81   929.96 12608.83
> sd(df$x48)
[1] 675.0915
> summary(df$x49)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  22.28  191.04  389.51  539.49  726.79 8989.37
> sd(df$x49)
[1] 529.7339
```

Question 1:

```
        Pearson's product-moment correlation

data:  df$x33 and df$x36
t = 23.556, df = 1262, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5131392 0.5898066
sample estimates:
     cor
0.552641
```

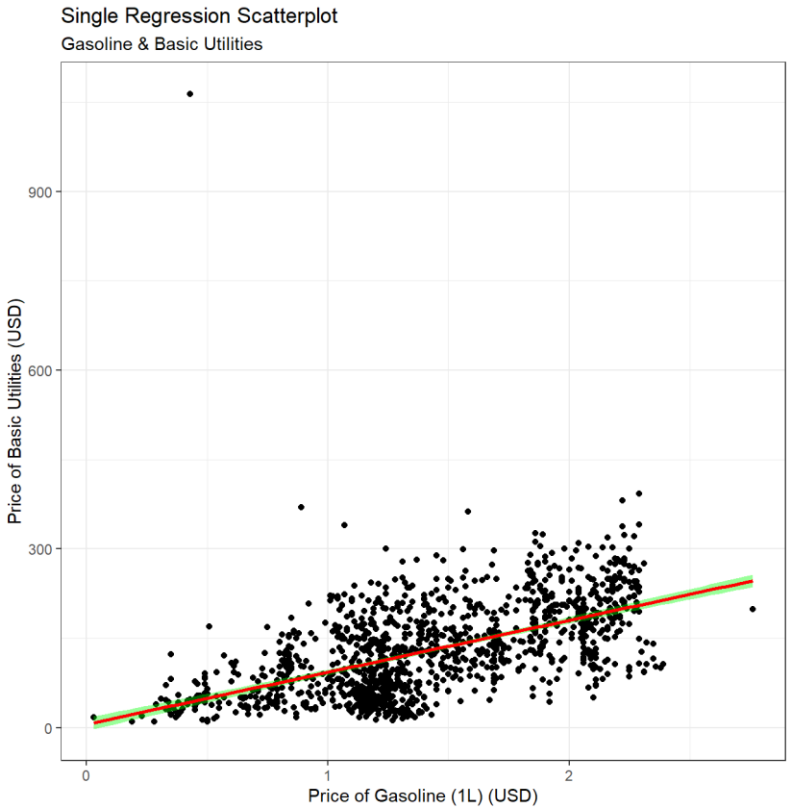
```
[1] "Reject the null hypothesis. We have enough statistical evidence to conclude the correlation is dif
ferent from 0."
```

```
Call:
lm(formula = df$x36 ~ df$x33)

Residuals:
    Min      1Q  Median      3Q     Max
-138.32  -44.08   -6.98   36.56 1020.20

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.899      5.573   1.058     0.29
df$x33        87.271      3.705  23.556   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 65.71 on 1262 degrees of freedom
Multiple R-squared:  0.3054,    Adjusted R-squared:  0.3049
F-statistic: 554.9 on 1 and 1262 DF,  p-value: < 2.2e-16
```



Question 2:

```
        Pearson's product-moment correlation

data:  df$x54 and df$x42
t = 43.147, df = 1262, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7487368 0.7933703
sample estimates:
      cor
0.7720035
```

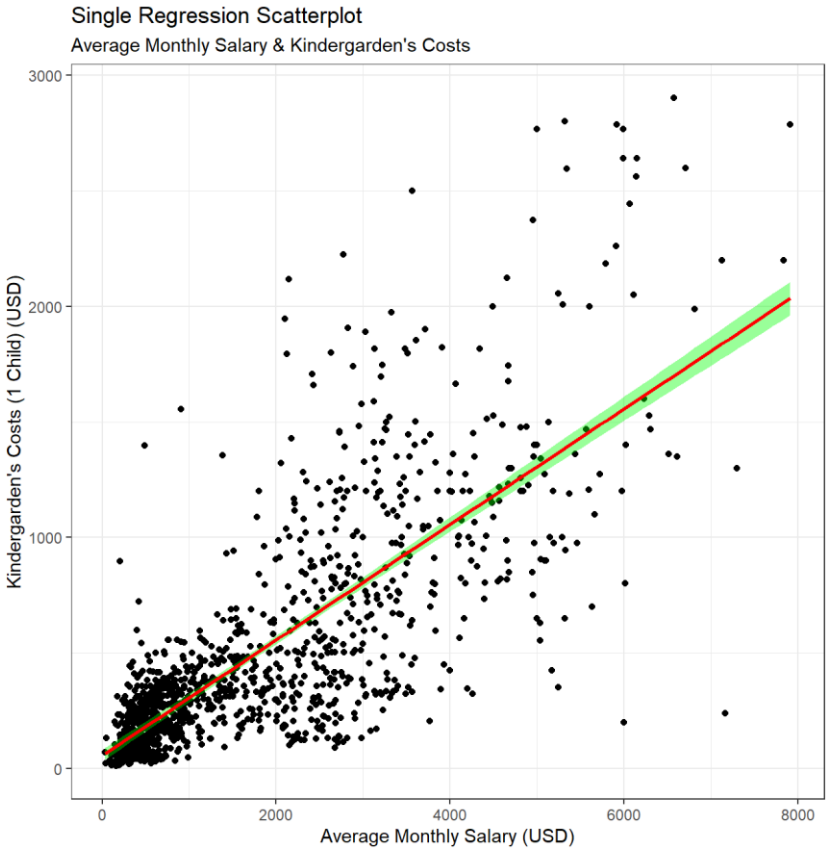
```
[1] "Reject the null hypothesis. We have enough statistical evidence to conclude the correlation is dif
ferent from 0."
```

```
Call:
lm(formula = df$x42 ~ df$x54)

Residuals:
     Min       1Q   Median       3Q      Max
-1610.22  -121.33   -23.88   101.93  1551.62

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 56.095800  13.321010   4.211 2.72e-05 ***
df$x54       0.250174   0.005798  43.147  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 314.4 on 1262 degrees of freedom
Multiple R-squared:  0.596,     Adjusted R-squared:  0.5957
F-statistic:  1862 on 1 and 1262 DF,  p-value: < 2.2e-16
```



Question 3:

```
        Pearson's product-moment correlation

data:  df$x49 and df$x48
t = 147.89, df = 1262, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9691627 0.9751959
sample estimates:
     cor
0.972341
```

```
[1] "Reject the null hypothesis. We have enough statistical evidence to conclude the correlation is dif
ferent from 0."
```

```
Call:
lm(formula = df$x48 ~ df$x49)

Residuals:
     Min       1Q   Median       3Q      Max
-1099.29   -50.98    -8.13    39.72  1482.50

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.300363   6.333893   5.257 1.71e-07 ***
df$x49       1.239149   0.008379 147.890  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 157.7 on 1262 degrees of freedom
Multiple R-squared:  0.9454,    Adjusted R-squared:  0.9454
F-statistic: 2.187e+04 on 1 and 1262 DF,  p-value: < 2.2e-16
```